AGILE: Autonomous Global Integrated Language Exploitation

Final Report (Year 2)

Contractor: **BBN Technologies**
10 Moulton Street
Cambridge, MA 02138

Principal Investigator: **Dr. John Makhoul**
Tel: 617-873-3332
Fax: 617-873-2473
Email: makhoul@bbn.com

Reporting Period: **1 November 2006 – 30 April 2008**

This material is based upon work supported by the
Defense Advanced Research Projects Agency DARPA/IPTO
AGILE: Autonomous Global Integrated Language Exploitation
ARPA Order No.: V002
Program Code No.: 5M30
Issued by DARPA/CMO under Contract #HR0011-06-C-0022

DEFENSE TECHNICAL INFORMATION CENTER

Information for the Defense Community

DTIC® has determined on Month: 05 Day: 15 Year: 2008 *that this Technical Document has the Distribution Statement checked below. The current distribution for this document can be found in the DTIC® Technical Report Database.*

☒ **DISTRIBUTION STATEMENT A.** Approved for public release; distribution is unlimited.

☐ **© COPYRIGHTED.** U.S. Government or Federal Rights License. All other rights and uses except those permitted by copyright law are reserved by the copyright owner.

☐ **DISTRIBUTION STATEMENT B.** Distribution authorized to U.S. Government agencies only. Other requests for this document shall be referred to controlling office.

☐ **DISTRIBUTION STATEMENT C.** Distribution authorized to U.S. Government Agencies and their contractors. Other requests for this document shall be referred to controlling office.

☐ **DISTRIBUTION STATEMENT D.** Distribution authorized to the Department of Defense and U.S. DoD contractors only. Other requests shall be referred to controlling office.

☐ **DISTRIBUTION STATEMENT E.** Distribution authorized to DoD Components only. Other requests shall be referred to controlling office.

☐ **DISTRIBUTION STATEMENT F.** Further dissemination only as directed by controlling office or higher DoD authority.

Distribution Statement F is also used when a document does not contain a distribution statement and no distribution statement can be determined.

☐ **DISTRIBUTION STATEMENT X.** Distribution authorized to U.S. Government Agencies and private individuals or enterprises eligible to obtain export-controlled technical data in accordance with DoDD 5230.25.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ORGANIZATION.

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
4/29/08	Final Report	1 November 2006 – 30 April 2008

4. TITLE AND SUBTITLE

DARPA/IPTO Final Report Option I/Phase II

5a. CONTRACT NUMBER

HR0011-06-C-0022

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

CLIN 000403

6. AUTHOR(S)

Dr. John Makhoul

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

BBN Technologies
10 Moulton Street
Cambridge MA 02138

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Defense Advanced Research Projects Agency DARPA/IPTO
3701 North Fairfax Drive
Arlington VA 22203

Dr. Joseph Olive

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

TBD

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Final Report

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Diane Messuri
U	U	U			19b. TELEPHONE NUMBER *(Include area code)* 617-873-2449

1 Introduction

The objective of GALE is nothing short of solving the human language technology (HLT) problem of transcribing foreign spoken and written languages into English and distilling the transcription into accurate information for use by our military. Below, we summarize the work performed by the BBN-led AGILE Team in Year 2. A more detailed description of the work performed can be found in the DARPA/IPTO Quarterly Status Reports for this project.

2 GALE Program Go/No-Go Targets

Starting with Phase 2 of the GALE program, the go/no-go targets for years 2 through 5 were changed. Table 1 gives the new targets for translation from text and transcription from audio (speech) for Phase 2. In the table, *consistency* is defined as the percentage of documents that must have at least the translation accuracy shown in the first column. As is clear from the table, these targets are different for the two languages (Arabic and Chinese), for structured and unstructured documents, and for text and audio input. For example, for structured Arabic text, the target is a consistency of 90% at an accuracy of 80%, which means that 90% of the documents must have a translation accuracy of at least 80%.

			Target	
			Accuracy	**Consistency**
Arabic	Structured	Text	80	90
		Audio	75	80
	Unstructured	Text	70	75
		Audio	70	70
Chinese	Structured	Text	75	90
		Audio	70	75
	Unstructured	Text	70	75
		Audio	65	70

Table 1: Targets for Phase 2 of the GALE program.

During the Phase 2 evaluations that were held in June 2007, the BBN AGILE team met or exceeded all four targets for Arabic, and the BBN team was the only team to do so. Not all targets were met for Chinese during the evaluation in June 2007. However, during the retest that was held in December 2007, the BBN team met or exceeded all four targets for Chinese. Again, it was the only team to do so.

In the distillation evaluations in June 2007, the BBN AGILE system had the best score in 9 of 12 conditions and had the best overall score. The system passed 7 out of the 12 targets and that was the best performance in the program. When compared to humans performing the same tasks under time-limited conditions, the AGILE system had higher recall (0.39 for AGILE compared to 0.37 for humans) but significantly lower precision (0.26 for AGILE compared to 0.45 for humans). However, both figures were better for the AGILE system than for the other two GALE systems in the evaluation.

3 Accomplishments in Speech-to-Text (STT)

3.1 BBN Technologies

Arabic STT

In the second year of the program, we continued to improve the performance of our Arabic STT system and achieved about 22% relative reduction in word error rate (WER) in comparison to the first year's system. This significant improvement was obtained through the implementation of multiple complementary systems for combination, dual audio segmentations to handle the mixed formal read news and casual spontaneous conversations found in broadcast speech, and through the use of additional training data. In addition to the primary phonetic system, we developed a graphemic system and a morphemic system. Overall, all three systems performed at a similar level individually but they seemed to be complementary to one another and produced substantial WER reduction through system combination.

The ROVER combination of all Arabic STT systems developed at BBN, Cambridge University, and LIMSI for the 2007 Evaluation benchmark test produced the lowest overall WER of 11.8% (9.9% for Broadcast News and 13.7% for Broadcast Conversations) among the results produced by the three GALE teams participating in that evaluation.

We also constructed a small list of affixes found in various dialects of Arabic to be added to the list of affixes used in the Buckwalter morphological analyzer so that we could derive pronunciations for non-MSA words. The addition did help in reducing the out-of-vocabulary rates for our phonetic and morphemic systems.

Mandarin STT

We also achieved significant reductions in character error rate (CER) in our Mandarin STT systems used in the 2007 Evaluation benchmark test in June 2007 as well as in the Chinese retest carried out in January 2008. In comparison to the 2006 Mandarin STT system, the 2007 system achieved 25% relative reduction in CER. The improvement was obtained through the use of additional training data, the implementation of a better pitch feature extraction algorithm, and the development of multiple complementary systems and dual audio segmentations. The new pitch feature extraction algorithm alone provided about 8% relative improvement. The ROVER system combination of three complementary systems using different acoustic models and the dual audio segmentations provided a modest 5% relative improvement over the primary single system. The rest of the improvement was due to the use of more training data.

Region-Dependent Transforms (RDT)

We continued to develop the RDT technique that was developed at BBN and used it in two alternative Mandarin STT systems for both the 2007 Evaluation and the retest. RDT is a feature transformation method in which a discriminative training criterion is used to optimize a set of linear projections with each projection for each cluster of HMM states. Even though the RDT Mandarin STT systems were only slightly better than the primary system, they seemed to contribute quite well when being used to combine with the primary system.

Bayesian Adaptation

We also continued to develop a Bayesian framework for adaptive training. Under this framework, we use a distribution of transforms rather than a point estimate. The likelihood of an utterance is computed as a weighted sum of the likelihoods obtained by transforming its features based on each of the transforms in the mixture, with the weights set to the transform priors. Experimental results on Arabic broadcast news exhibited increased likelihood on acoustic training data and improved recognition performance on unseen test data, compared to speaker independent and standard adaptive models.

STT/Machine Translation (MT) Integration

Translating Confusion Networks Created from STT Output

We implemented software tools that enable us to decode lattice-style superposition of STT hypotheses, rather than decoding a single-best STT hypothesis. The purpose was to transition towards tighter integration of STT models into MT, in a way that would allow us to directly tune system components to minimize translation error rate. We conducted experiments measuring the effect of various system improvements on the performance of our confusion network decoding system, and we measured the difference in performance between confusion-network decoding and simply decoding the 1-best STT hypothesis. We observed small improvements in performance over the baseline.

Optimizing STT System Combination for MT

We implemented an optimization pipeline for tuning the parameters of system combination for STT. We are now able to tune the STT combination weights and MT decoding weights jointly, using the recently implemented consensus network translation capability. Our preliminary results show that it is possible to achieve slightly improved MT performance from this type of tuning, but the gains are not consistent across all test sets.

3.2 Cambridge University

Arabic STT System

A complete high-performance Arabic STT system was built and used as part of the AGILE Arabic 2007 STT system (via ROVER combination with systems from BBN and LIMSI). The Cambridge system uses multiple branches and includes both phonetic acoustic models which tend to be more accurate for MSA data; and graphemic models which allow higher lexical coverage and are more robust to pronunciation variation. The combined system uses cross-adaptation from graphemic to phonetic models, multiple acoustic segmentations, along with both Gaussianized and standard front analyses.

Improved Phonetic Models for Arabic STT

We developed a method for estimating a set of pronunciations for unknown words which allows the lexical coverage of a phonetic Arabic STT system to be greatly increased.

Class Based and Sub-word Language Models for Arabic STT

Due to the productive morphology of Arabic, there are a very large number of word forms observed in Arabic texts. Hence the Arabic language models are very sparse and OOV rates tend to be high. Automatically derived class-based language models have

been investigated for Arabic for vocabularies up to 350k words and have led to reductions in error rates. Furthermore sub-word models, including morpheme-level and automatically derived sub-word models have also been investigated with the aim of reducing sparseness. Work is continuing in these areas.

Improved Chinese STT System

The performance of the Cambridge Chinese STT system was improved by about 9% relative by tuning the vocabulary and including more training data. The AGILE Chinese STT system used the BBN system for adaptation supervision for the Cambridge system which generates the final output. This form of combination was optimized for translation performance but also reduced speech recognition error rates by about 10% relative over the year. The Cambridge system without BBN adaptation input was used to process Chinese audio for distillation.

Unsupervised and Directed Manual Transcription

We have studied the use of unsupervised acoustic model training for Chinese broadcast conversation and broadcast news recognition. The selection of particular blocks of automatically transcribed data, via confidence scores at the segment and show level, has been investigated. Furthermore the selective manual transcription of data that are particularly problematic has been tested. Using this approach, a large proportion of the error rate reduction from new data can be obtained while only providing manual transcripts for a relatively small data fraction.

Advanced Acoustic Modeling

A number of advanced techniques for acoustic modeling have been investigated. These include a new method of discriminative estimation of adaptation transforms known as discriminative mapping transforms, semi-parametric trajectory modeling which is an extension of the fMPE technique, the use of directed decision trees, and Bayesian adaptive training.

Improved Sentence Segmentation

The acoustic speaker-turn segmentation used as part of the STT system needs to be refined to provide sentence-level segmentation for machine translation. Previously we had used a hidden-event language model and a simple heuristic based on pause durations for the sentence boundary detection task. This was improved to include a decision-tree based prosodic feature model integrated with the language model. The new structure shows improvements in translation of up to 1.9% TER. A number of further improvements in the feature set and structure of the prosodic feature model, including the use of support vector machines, has been investigated.

Discriminative Language Model Adaptation

The N-gram language models for STT are an interpolation of a number of component language models from a number of sources/genres.

Unsupervised language model adaptation is performed to tune the language model for a particular story by varying the interpolation weights based on an initial transcription with a non-adaptive system. These weights are either adapted to minimize perplexity or to minimize the expected STT error rate. Both methods have been found to yield small

consistent gains in error rate and a similar approach has been applied to translation language models and has also yielded improvements.

Chinese Retest

The work for the Chinese retest involved continuing work in improving Chinese STT; rebuilding and running the Cambridge/AGILE STT systems for the retest evaluation itself; applying MT improvements previously developed to Chinese and producing system outputs for all Chinese genres for the retest data; and finally investigating various segmentation effects for STT-MT integration.

After the retest evaluation there was effort in error analysis and presentations on the work done for the retest evaluation.

Work on Chinese STT continued on from the work done in Phase 2 readying systems for the Chinese retest. Substantial reductions in error rate had been obtained by improved pitch features, and the use of revised build procedures leading to larger language models. These new processing were used to rebuild STT systems including more data for acoustic and language model training. Furthermore an additional re-scoring stage using the fMPE technique was also included in the system which yielded a small benefit. The overall AGILE Chinese STT system (in which the Cambridge system is cross-adapted to outputs supplied by BBN) was used for the retest. Overall the cross-adapted Cambridge AGILE Chinese STT system reduced the character error rate by 17% relative from the June'07 evaluation system, roughly half of this was due to improved modeling and half due to manual segmentation. An investigation of further automatic segmentation applied to the given manual segmentation was also carried out, although this yielded mixed results and was not used in the Chinese retest system itself. After the retest evaluation was run, an error analysis error was performed and more experiments performed on the effects of manual segmentation on STT performance and also on alternative methods of STT hypothesis combination.

New translation systems were built for all Chinese genres based on revised training sets which overall yield up to 2% absolute improvements in the BLEU translation measure. A zero cut-off 5-gram model for Chinese has also been used, and a revised minimum Bayes risk decoding procedure investigated. These systems were included in the MT system combination used in the AGILE Chinese retest evaluation.

3.3 Georgia Institute of Technology

A new discriminative training method, called soft margin estimation (SME), was proposed for estimating the parameters of continuous density hidden Markov models. The proposed method makes direct use of the successful ideas of soft margin in support vector machines to improve generalization capability and decision feedback learning in discriminative training to enhance model separation in classifier design. The proposed algorithm has already shown its advantage over traditional discriminative training methods (such as MCE) in several speech recognition tasks.

In the connected digit TIDIGITS task, the string accuracy using 12-state 16-mixture models was 99.30%. By jointly optimizing the acoustic feature and HMM parameters, a 99.61% string accuracy was obtained. Even for 1-mixture model we can achieve a string accuracy of 99.13%.

For the medium vocabulary continuous speech recognition (WSJ) task, we treated SME thoroughly with extended Baum-Welch (EBW) optimization. SME realizations based on utterance selection and frame selection were realized. In contrast to SME with utterance selection, SME with frame selection uses more confusion patterns. For a fair comparison with MCE, the implementation of SME and MCE shares the same core components. Tested on the 5k-WSJ0 task, all the proposed SME methods achieved about more than 12% relative WER reductions over the MLE baseline (5.06% WER). Moreover, all SME methods converged in 15 iterations and outperformed MCE (4.60% WER). The SME model with frame selection achieved 4.11% WER, with 19% relative WER reduction from MLE, and 10% more relative WER reduction than MCE.

As SME is now optimized with EBW and performs better than MCE, current research work builds a good basis for SME to be used in LVCSR tasks in future as an alternative discriminative training method. Our preliminary results showed SME improvement over MPE.

3.4 LIMSI

The main LIMSI activities in the second year of the GALE program have addressed improving Arabic STT with a small effort devoted to Mandarin STT. The work has addressed methods to facilitate training with incomplete information and to improve pronunciation modeling; to incorporate prosodic features (duration, pitch); updating the connectionist language models; and investigating the use of multi-layer perceptrons (MLP) to estimate long span features.

The majority of Arabic transcriptions are not vocalized. Thus, one of the costly steps in acoustic model training is extending the pronunciation dictionary to cover any additional words. For many languages this is relatively straightforward via grapheme-to-phoneme conversion. In Mandarin this does not pose much of a problem since the new words can be split into their constituent characters. For Arabic, the problem reduces to that of determining the vocalized form, after which grapheme-to-phoneme conversion is relatively simple. The Buckwalter morphological analyzer is generally used to all possible vocalizations of a word, but there are some words for which no forms are able to be derived. Some simple rules were developed that are applied in pre- and post-processing to enable Buckwalter to produce more derivations. These rules decompose the words into one or more word sequences and look up the constituent parts in the Master dictionary or apply Buckwalter to them. The Buckwalter prefix table was augmented to include affixes found in Arabic dialects, thus permitting some dialect words to be vocalized. A limited set of vocalization with a 'generic' vowel are generated for words that could not be handled by the extended Buckwalter analysis. The generic vowel serves as a place holder thereby substantially limiting the number of possible combinations compared to three vowels. Of course, such automatic processes typically do not generate pronunciation variants, the study of which is also underway. Other pronunciation model research has reconsidered the phone set, introducing explicit symbols for geminate consonants and the tanwin. The estimated improvement that can be attributed to the above work is 0.5% absolute.

It is well known that HMMs do not properly model the phone and the word durations. The segment duration being implicitly encoded in the model topology, the transition

probabilities, and the derivative features, none of these model parameters can properly capture segment duration when considering a wider context than a triphone. A new strategy was proposed to add duration information as a post-processing of the decoding process, via a word lattice representation which also includes the phone segmentation for each word edge is used. This approach allows duration information to be used in conjunction with consensus decoding. The duration model gives small, but regular gains of 0.2-0.4% absolute gain on the GALE Arabic data.

Another activity has aimed at improving the language models and in particular the continuous space language model. The basic idea of such a model is to convert the word indices to a continuous representation and to use a probability estimator operating on this space. Since the resulting probability functions are smooth functions of the word representation, better generalization can be expected. The training method has been improved to make use of a resampling algorithm, and training multiple connectionist LMs with different configurations. The connectionist LM is then interpolated with a standard 4-gram back-off LM, word error rate reductions on the order of 0.7 to 0.9% absolute are obtained compared to the 4-gram back-off LM.

Effort was also allocated to carrying out an analysis of the STT errors. Roughly 15% of the errors involve affixes, which are more often deleted than inserted (ratio 2:1). One third of the prefix errors involve the article 'Al'. As a result of this analysis, the pronunciation dictionary was extended to add alternate pronunciations for the final a/e variant observed in Lebanese dialects.

MLP features have been attracting growing interest for LVCSR due to their complementarity with cepstral features, their object being to augment the short-term spectral representation provided by PLP or MFCC features with more contextual information. Different types of MLP features and their combination have been explored, and evaluated both without and with unsupervised model adaptation. Extending the feature vector by combining various feature sets (one based on nine frames of PLP features and their derivatives with a temporal span of 150 ms, and the other an improve LP-TRAP that has a longer temporal span of 500 ms) led to a 9% relative WER reduction relative to the PLP baseline. Significant gains are also reported with both ROVER and cross-model adaptation. Feature concatenation appears to be the most efficient combination method, providing the best gain with the lowest decoding cost. In general, it seems best to combine features based on different time spans as they provide high complementarity.

Since much of the algorithmic and methodology advances are language independent, a small effort has been allocated to improving our Mandarin STT system. The character error rate has been reduced by 10-15% relative, depending upon the data set. The development of pitch features has been explored for Mandarin and we are in the process of incorporating these in the Arabic system.

Prior to the June 2007 evaluation, the decoding strategy was completely revised, and the acoustic and language models improved. The main changes are: the use of pronunciation probabilities during lattice generation; the use of duration models in all decoding passes; using the best acoustic models in the first decoding pass (previous systems used smaller models); 3 decoding passes; and MLLR adaptation with a 2 regression classes so as to get

better confidence scores. Compared with LIMSI component used in the June 2006 AGILE Arabic STT system, the word error rate has been reduced by almost 20%.

4 Accomplishments in Machine Translation (MT)

4.1 BBN Technologies

Data Processing for MT

BBN took charge of all data processing tasks for the AGILE team, developing procedures for data tokenization and normalization, sentence alignment, segment filtering, etc. Several revisions of this data pipeline were performed in order to correct found problems and to incorporate new data released by LDC on a regular basis. The largest gain (about 2 BLEU points) from incorporating new data came from adding Sakhr's Arabic-to-English and English-to-Arabic parallel training data.

System Combination for MT

We continued to improve the AGILE MT system combination procedure, adding the following capabilities:

- proper use of word confidence scores derived from system N-best outputs
- automatic estimation of the prior probability that a system output can serve as the "skeleton" for aligning other MT outputs to it
- rescoring of the final N-best generated from the confusion network using a large, un-pruned 5-gram language model (LM)
- integration of a bi-gram LM in the decoding of the confusion network for N-best generation, as well as iterative optimization of system weights
- improved construction of the confusion networks by incrementally aligning each system output to the skeleton hypothesis.

The above enhancements resulted in over 2 BLEU points improvement in MT accuracy. We also explored an alternative method for combining the BBN and Sakhr MT outputs, in which the Sakhr MT output is used to determine the overall structure of the translation and the BBN system is used to provide alternative translations for the noun phrases (NP) occurring in the sentence. This constrained form of system combination leads to significant improvement in TER/BLEU (over 10 points) compared to the standalone Sakhr system. We are currently investigating the effect on human-mediated TER (HTER).

Translation of Special Entities in Chinese

We devoted attention to the translation of special entities. Handling of numbers, dates, and times was improved by enhancing the tokenization process to identify such sequences in the input text and map them directly to the corresponding English form. Proper translation of Chinese names to English was more involved, requiring the use of a name translation list that was automatically extracted from the Chinese-English parallel training data. We also implemented rules for the translation of Chinese bylines, which frequently appear in newswire text. This collection of rule-based translation techniques resulted in significant improvement of translation quality for names, numbers, dates, times, and bylines.

Error Analysis

We performed a thorough error analysis of Chinese MT. Major categories of errors are 1) deleted content words, 2) incorrect sentence and phrase structures, 3) missing information in the source, 4) word sense ambiguity, and 5) word segmentation errors, particularly segmentation of proper names. Many of these are also true for Arabic MT.

Translating with Multiple Chinese Word Segmentations

Chinese character-to-word segmentation is ambiguous, leading to translation errors in many cases. To alleviate this problem, we have implemented a decoding process in which multiple word segmentations are considered simultaneously. This technique provided about 0.3 BLEU gain in translation accuracy, and it also gave translations that seemed complementary to those of the regular word segmentation, thus helping improve the final AGILE system combination for the Chinese retest.

Discriminative Corpus Weight Estimation

We have begun to investigate a novel approach for automatic training data filtering, in which each corpus (or collection of sentences) in the parallel training data is assigned a weight that reflects its importance (in terms of contribution to machine translation accuracy) in the estimation of translation rule probabilities. The corpus weights are optimized based on a gradient descent method so as to minimize translation edit rate (TER) on a held out set. So far, only modest improvements in TER/BLEU have been obtained by using discriminative corpus weights, but the method is still under development.

Improved Arabic MT through Morphological Analysis

Arabic is a morphologically rich language, where each word may consist of a prefix, stem and suffix. The combination of stems and affixes leads to a very large vocabulary, so regular hierarchical or phrase-based translation becomes problematic due to the sparse space of word n-grams. The out-of-vocabulary (OOV) rate is also high on documents that are slightly out of domain with respect to the parallel training data (e.g., web logs or newsgroup text). To alleviate this problem, we processed the training data so as to split Arabic words into their stems and affixes based on Sakhr's morphological analyzer, which parses the Arabic text and assigns morphosyntactic tags to each word. Such morphological decomposition helps reduce the OOV rate on unseen data and results in more robust translation rules, providing up to 2 BLEU points gain in MT accuracy.

Improved GIZA Training

We improved GIZA alignments by segregating parallel data into high-quality and low-quality sets, then splitting the low-quality data into subsets and adding the high-quality data into each subset prior to running GIZA. This training procedure led to significant improvements in performance on Arabic (about 1 point gain in both TER and BLEU), where the parallel data obtained from Sakhr was used as the high-quality set. We also successfully applied this procedure to Chinese, where the high-quality set consisted of newswire, GALE corpora, and lexicons.

Percentile-based Optimization for MT

At the Phase 2 kickoff meeting, it was announced that the target metric would be changed from an average performance to a minimum accuracy of a certain percentage of documents. We, therefore, decided to investigate whether it is possible to tune the system combination or the MT decoder weights so as to increase the 90th percentile document accuracy on automatic metrics. We implemented changes in the automatic optimizer to allow tuning the various weights to optimize different functions. We found that the gains from this optimization were very small and typically did not carry over to new test sets.

Understanding-empowered MT

We designed and developed a powerful hierarchical machine translation engine that allows easy experimentation with new features and techniques. Taking advantage of this flexibility, we made a first step towards our goal of understanding-empowered MT by implementing string-to-tree translation, where the input source sentence is translated into an English dependency tree. Through this method, the set of hierarchical translation rules was reduced by 80% with no degradation in MT accuracy. Integration of an English dependency tree LM gave a gain of 1.2 in BLEU; Adding target-side POS tags in translation rules and penalizing long-distance re-orderings provided additional gains (up to 0.2 and 0.3 BLEU, respectively).

Genre-dependent Optimization for MT

We conducted a set of translation experiments using genre-specific MT systems to see whether we can improve performance on genres that are not well represented in the parallel training data, such as newsgroups/weblogs and broadcast conversational data. The results of these experiments show that genre-specific modeling can provide significant gains in translation accuracy on newsgroups (up to 1% gain in TER and BLEU).

Dialect-MSA Mapping for Arabic-English Translation

We applied dialect-to-MSA word mappings on the output of STT before running the MT. The mappings are at the phonetic, morphological and lexical levels, providing a modest improvement of 0.15 to 0.2 points in TER for BC.

Language Modeling for MT

Given earlier indications that larger LMs in decoding can help improve MT accuracy, we investigated several ways to increase the power of the English LM: going beyond 5-gram LMs in rescoring, trying a part-of-speech (POS) LM of large n-gram order as an additional score, rescoring lattices of translations using un-pruned LM, and decoding with un-pruned LMs distributed over a number of compute servers. The latter technique gave significant improvement (about 1-2 BLEU points), while the other methods provided either too small or inconsistent gains.

4.2 Cambridge University

Morphosyntax-based Arabic Word Tokenization

We have investigated the incorporation of Arabic morphological analysis into a phrase-based translation system. MADA, a full morphological tagger for Modern Standard

Arabic developed at Columbia University, takes decisions on the best analysis for each word, given the alternative analyses provided by the morphosyntactic models from the Buckwalter analyzer. Using this analysis of the Arabic parallel text (AGILE v3.1), the number of distinct tokens is reduced from 612.4K to 421.6K. We found substantial improvements in translation perform both from text and from STT output; the process appears to be robust to STT errors. This analysis was used in the Cambridge component of the AGILE 2007 evaluation system.

Chinese-English Lattice Text Translation

Direct translation of Chinese word lattices was used for two different translation problems. Reordered Chinese text mapped into English word order by MIT was added to Chinese in its 'natural' as the first step in translation. This analysis was used in the Cambridge component of the AGILE 2007 evaluation system. Direct translation of STT lattices was also investigated with the aim of introducing alternative word hypotheses into translation, with small but consistent gains resulting in BLEU and TER.

Minimum Bayes Risk Decoding for SMT

A standard approach used in minimum Bayes risk decoding for SMT systems is to compute an expected sentence level cost function. We investigated balancing TER and BLEU optimization through variations of the loss function used in MBR decoding. Relative to the maximum likelihood translation hypothesis, we found significant TER improvements from the improved MBR-BLEU approach, as well as significant changes in the top hypotheses in the re-ordered N-best output. By applying MBR in this way, a send a second Cambridge component MT evaluation system was produced.

MT System Combination

Multiple confusion network and the associated hypothesis selection schemes were investigated for MT system combination. Strategies were studied to ensure that the hypothesis used as the alignment skeleton was produced by a system with relatively high system weight. By appropriately selecting the hypothesis used to guide the system combination, significant reductions in TER were produced, especially in the translation of Arabic and Chinese audio.

Zero-cutoff 5-gram Language Models in Translation

Recent work by Google has demonstrated that long-span language models can be applied without any form of backoff when very large amounts of language model training data are available. We are developing efficient estimation and implementation strategies for these models and our initial experiments have yielded significant improvements in BLEU and TER for translation of Arabic newswire and webtext.

Phrase-based Language Models in Translation

We have developed modeling and estimation procedures for phrase segmentation models. These can be considered loosely to be 'phrase n-grams', although they are more exactly probabilistic transducers which assign likelihood to sequences of English words generated as translation hypotheses. These models were motivated by our interest in identifying aspects of the translation systems whose components could be estimated over monolingual (English) text. We find that estimating phrase segmentation transducers

over monolingual English gives complementary translation performance gains to improvements in the word-based language models, although the estimation strategies which are most effective for phrase sequences are not those which work best for words due to the very different distribution of phrases.

Translation System Development

Following the GALE 2007 MT Evaluation, we reimplemented the translation system to make use of the recently released OpenFst toolkit. Work is ongoing; however we have so far seen extensive reductions in processing time and memory usage, mainly through improved pruning strategies, with no loss in translation performance.

Participation in the GALE 2007 Machine Translation Evaluation

Cambridge submitted systems in all conditions for the Arabic and Chinese to English translation evaluation.

4.3 Information Sciences Institute (USC)

In the first three quarters of Phase 2, ISI obtained these BLEU score improvements for its syntax-based translation systems on AGILE 4.1 test sets:

- Chinese/English -- from 39.7 to 45.1
- Arabic/English -- from 51.9 to 53.2

These improvements came largely from the following areas of research.

Syntax Structure and Alignment

Our system learns syntactic translation rules from <English tree, foreign string, alignment> triples in the training data. To create such triples from bilingual corpora, our baseline is to apply GIZA word-alignment models, and to automatically parse the English half of the training data. In Phase 2, we improved significantly on this baseline in five ways:

- We employed a new generative model for word alignment, called LEAF.
- We re-structured the English parses. Binarizing flat structures (especially noun phrases) leads to more general translation rules. Overall translation accuracy improves +0.5 points when flat structures are binarized to the left, but a full +1.0 point when we binarize adaptively with EM.
- We re-aligned the data using syntax. After we obtain a syntax translation model from re-structured trees, we throw away the LEAF/GIZA word alignments and employ EM to re-align by optimizing the parameters of the syntax model. This yields +1.0 BLEU for Chinese/English and +0.5 BLEU for Arabic/English.
- We re-labeled the English parses. We trace many translation errors to overloading of syntax symbols such as VP (verb phrase) -- a low-level rule may supply a VP to a high-level rule, but it may be the wrong kind of VP. When we split categories in a controlled way, we obtain a +0.3 BLEU for Chinese/English.
- We trained our parser on more Treebank data. Moving from 1m words of Treebank to 2m words yields a +0.5 BLEU improvement for Chinese/English.

Decoding

Until recently, decoding with syntax-based models was impractical due to massive search errors. We overcame this barrier and improved the functionality of our decoder in these ways:

- We integrated 5-gram language models into decoder search. This gives a gain of +1.4 BLEU points for Chinese/English over the previous 3-gram language model.
- We reduced our observed decoding time from n^2.7 to n^2.2 (where n = input sentence length).
- We provided support for multiple language models, lattice-input decoding, and duplicate-free n-best lists.
- We designed a scalable rule database and used 500m learned syntax translation rules in the GALE-07 evaluation.

Compact Language Models

We developed a way to use minimal perfect hash functions to compress (by a factor of 10) the space needed to store a language model, at the cost of bounded errors in probabilities returned. Decoding experiments confirmed no BLEU loss.

Name Translation

We built a novel Arabic tagger that automatically identifies words that should be transliterated into English. We improved the f-measure of this tagger from 72.0% to 86.2%. Words so tagged are processed by our Arabic/English transliterator module. Name, number, and date modules contribute a total of +0.7 BLEU to the Arabic/English system.

In the last quarter of Phase 2, we began new, large initiatives in the following areas: (1) modeling inter-rule context, (2) scaling up to more syntactic features with perceptron training, (3) synchronous tree-insertion grammar, and (4) tree-based target language models. We expect these all to pay off in Phase 3.

4.4 Massachusetts Institute of Technology (MIT)

Reordering Approaches for Chinese

In Year 1 we developed reordering approaches for translation from Chinese to English. In this approach, the Chinese source text is first parsed, and rules are then applied to the resulting parse trees in an effort to transform the Chinese text into a word order that is closer to English. This pre-processing (reordering) step is applied to Chinese sentences in both training and test data. In Year 2 we made a number of improvements to the basic reordering approach. We improved the Chinese word segmentation algorithm. We investigated more reordering rules, and ran detailed evaluation experiments to assess their relative accuracy and efficacy. Our experiments showed encouraging gains with the MOSES phrase-based system. The reordering approach improved the BLEU score for the MOSES system from 28.52 to 30.86 on the NIST 2006 evaluation data. The reordering approach was used within the MIT and Cambridge phrase-based translation systems used in the GALE evaluation. This work resulted in an EMNLP publication (Wang, Collins, and Koehn, 2007), which describes the approach in detail.

A Perceptron-trained Model for Syntax-based Translation

A new focus of our work has been the development of a perceptron-based model for translation from Chinese to English, based on the approach described by Cowan, Kucerova, and Collins (EMNLP 2006). Our goal is to learn a model that maps parse trees in the source language to parse trees in the target language. The model is learned from a corpus of translation pairs, where each sentence in the source or target language has an associated parse tree that is recovered using a treebank-trained parser.

In our approach, a parse-tree structure in the target language is built step-by-step, based on information from a source-language parse tree. The steps involve choices such as the identity of the main verb, the argument structure of the verb (i.e., whether it has a subject, object, or other arguments), and alignment information specifying where constituents in the source-language tree should appear in the target language translation.

Our first step was to complete a major goal in training the model, which was to extract aligned syntactic structures from the training corpus. The result of this step was a set of training examples, where each training example consists of a pair of aligned syntactic structures. More specifically, each training example consists of a full Chinese clause structure paired with an English structure which contains: (a) the main verb in the English clause, (b) function words in the clause, such as complementisers, or auxiliary verbs, (c) alignment information, specifying where each modifier in the Chinese clause should be placed in the English clause. We have now trained a perceptron model on this data, which predicts English syntactic structures from Chinese input sentences. Current work is focusing on integrating this model within a full translation system.

4.5 Sakhr Software

Sakhr Software performed several tasks which were aimed at improving the overall performance of Arabic-to-English MT. The following is a summary of the tasks performed.

We translated GALE06 data with automatic segmentation using the latest Sakhr MT system. We also provided morphological tagging (1-best) of 1000 sentences from MT05 and 1-best translations of whole sentences with noun phrases marked in the source and target.

We provided a corpus of about 27 million source words, which is divided into two parts: an Arabic-to-English (A2E) part totaling about 19 million words, and an English-to-Arabic (E2A) part totaling about 8 million words. The translations were obtained by first running the data through the Sakhr machine translation system, then correcting the output by hand. We have also morphologically tagged this corpus.

We have been able to design and implement a slang-to-MSA converter which will feed into the Sakhr MT system to obtain an Arabic slang-to-English translation. The design consists of two modules, a dictionary of slang expressions and their corresponding MSA equivalents, and a rules engine that detects the occurrences of slang in the input and converts them into MSA. This was accomplished by adding 6,000 n-grams (bi/tri/quad) and about 30 rules.

We implemented a generator for new English idioms by using overlapped (existing) idioms. This enables auto expansion of the Arabic monolingual lexicon and the Arabic-to-English Transfer Lexicon.

We have exposed disambiguated morphological tags in the output of the MT. These were used in the training of BBN SMT engine, to split affixes and diminish sparseness in the translation model.

We adjusted the morphological tagger to handle single-word sentences more efficiently. This leads to better overall translations when single-word sentences occur.

We integrated the overlapped idiom module into the main MT engine. The result is better English fluency in the translated output.

We started work on the generation of multiple translations. The first phase is a system combination, where Sakhr's translation was the basic translation and selected noun phrases of the source sentence were translated through the BBN engine. The second phase will expose various decision points internal to the MT engine, which will facilitate the combination with a statistical engine leading to a true hybrid MT system.

4.6 University of Edinburgh

Factored Models

We continued work on factored translation models, which allow the integration of additional annotation at the word level by representing words as vector of factors, instead of simple tokens.

In an extension of the basic approach that we call "asynchronous factored models" we eliminate the constraint that all factors must be translated at the same time, and that source and target lengths for those factors are identical for each translation steps. This allows us to decode larger phrases for factors which are less sparse where we would be confident of longer translated phrases, for instance utilizing part-of-speech tags for reordering of larger segments.

Factored translation models allow a model using a decoding path that first translates lemma and syntactic information separately and then generates surface forms. Backing off to the translation of lemmas instead of surface word forms should enable a more robust transfer step. However, in practice, there is a loss when throwing away surface form information (and information about translating surface forms). To gain benefits from a more robust model, while not discarding knowledge about surface forms, we would like to combine both models in a so-called alternative path model, where phrase translations may be derived from any of multiple possible decoding paths.

Discriminative Training

Current state of the art statistical translation models consist of smaller, generatively trained component models (e.g. language model, translation model) which are treated as features of a log-linear model. Typically, the weights of the log-linear model are learned discriminatively by optimizing against a particular objective function (conditional probability, minimum error rate). In contrast to this, our approach seeks to train all model parameters discriminatively. Instead of having a dozen or so features, we employ millions of features.

We use a perceptron-based large margin learning algorithm to train our model. The model is optimized on a simple 0/1 sentence error loss function, or a loss function based on automatic error measures such as BLEU. Our model uses millions of features whose weights are learned discriminatively. Every entry of a pre-existing phrase table is treated as a feature. We are also able to learn a discriminative language model by treating target side n-grams as features. Other features include source-side, source side distortion lengths and reordered phrase pairs.

Bloom Filter Language Models

Larger training corpora for the language model promise better performance. Since large amounts of English text are easily available (billions and trillions of words), the main challenge is to handle the massive quantities of data. During machine translation decoding, the language model is consulted very frequently. Hence, the language model is best kept in RAM, preferably on a single machine.

We investigated the use of lossy data structures to store the language model, so-called Bloom Filters. A Bloom Filter represents a set $S = \{ x1, x2, \ldots, xn \}$ with n elements. The only significant storage used by a BF consists of a bit array of size m. This is initially set to hold zeroes. To train the filter we hash each item in the set k times using the distinct hash functions h1, h2, …, hk. Each function is assumed to be independent from each other and to map items in the universe to the range 1 to m uniformly at random. The k bits indexed by the hash values for each item are set to 1. Afterwards the item is discarded. Once a bit has been set to 1 it remains set for the lifetime of the filter. Distinct items may not be hashed to k distinct locations in the filter; we do not keep track of collisions. Locations in the filter can, therefore, be shared by distinct items allowing significant space savings but introducing a non-zero probability of false positives at test time. There is no way of directly retrieving or enumerating the items that have been entered into a BF.

We have implemented a first version of Bloom Filter language models and demonstrated that we are able to make better use of available RAM than traditional language models (better performance given same amount of RAM). In future work, we plan to extend the approach to integrate more sophisticated smoothing techniques, such as Kneser-Ney discounting.

Domain Adaptation

A well-known problem for machine translation is the application of machine translation system mainly built for one domain to a different domain. Within the GALE project such

domains are news wire, broadcast news, broadcast conversation, and web content, but most of the training data is from sources such as the United Nation reports.

We train multiple translation models and languages models, for instance trained on an in-domain data set and additional models trained on a general data set, typically consisting of the one training data. Phrase translations are taken from either translation table and scored with both language models. Weights for all the different models are set with minim error rate training.

Alternatively, we may cluster the training data automatically into, say, 10 clusters, and train separate domain models for each. By matching test data to any of the 10 clusters, we detect which domain they belong to and find the appropriate in-domain model.

4.7 University of Maryland

Optimization Using Paraphrased References

We developed techniques for parameter-tuning using a small set of human references that are expanded by means of automatic paraphrasing. We showed that when we only have a single human reference translation, these artificially expanded references could improve the TER and BLEU scores by about 1.5 points – almost as much as the gain resulting from using extra human references.

Class-Based Rule Generalization using Bilingual Classes

As a first step toward an unsupervised multi-class hierarchical model, we developed a clustering method for bilingual word classes. We demonstrated that class-based rule generalization based on these clusters improved translation over a baseline system by about 1 BLEU point. However, the gain was only slightly more than that achieved by using a single class for all bilingual word pairs.

Constrained Decode in the Hierarchical Decoder

We developed a constrained decoding capability within the BBN Hierdec hierarchical decoder. The decoder can take one or more required target strings and force the decoder to find the best possible translation score for each of these target answers. This capability is useful for diagnostic purposes and also for forcing two different models to evaluate the same possible answers so as to combine their scores.

Experimentation with Phrasal EM

We developed and evaluated a phrasal EM method for improving the alignment of source and target words, such that the likelihood of the training data under the translation model would increase. We made the constrained decode within the phrasal decoder very fast so that it would be possible to perform this function on a large training set. However, our experiments did not yield any improvement in the resulting translation accuracy over the common one-pass method of aligning using GIZA++ followed by phrase extraction.

Alignment Combination using Learning

We applied the maximum entropy alignment combination method, as described in (Ayan and Dorr, 2006) in the BBN system and also combined the unidirectional alignment from the ISI system with the bidirectional GIZA++ alignments.

This improved our MT systems significantly (up to 1 BLEU point on Arabic-English and 0.5 BLEU point on Chinese-English).

Human-Targeted Translation Edit Rate

We developed an objective evaluation measure that has higher correlation with the GALE program HTER measure by combining scores and features from several different automatic evaluation metrics using an artificial neural network. Experiments showed a small increase in prediction accuracy and correlation with HTER versus any single automatic evaluation metric.

Unsupervised Adaptation of the Translation Model

We implemented an unsupervised adaptation method in which we perform a first translation pass and then take the resulting output, together with the input, as additional training data, and finally perform a second translation pass. The improvement from this method was rather small – about one half point.

Improving MT using Comparable Corpora

We are working on methods for improving machine translation using comparable corpora. We use cross-lingual IR to find documents that share many words and concepts with the documents being translated and then attempt to learn new word and phrase translations from these comparable documents. Initial work has shown gains of 2 TER points and 1.5 BLEU points on older versions of the MT system. The method is being reevaluated on the latest version of the BBN Hierdec MT system.

5 Accomplishments in Distillation

5.1 BBN Technologies

In Phase 2 of GALE, the AGILE Distillation team adapted and improved our distillation system to address new genres, new templates, and new redundancy and citation requirements. We successfully participated in both the Phase 2 Utility Evaluation and the Phase 2 Go/No-Go Evaluation, ranking first in the majority of all evaluation categories. The AGILE Distillation system was also part of a successful demonstration of GALE technology at DARPATech 2007.

User Application

During the first half of Phase 2, the AGILE team undertook a major redesign of the skeletal user interface implemented in Phase 1, focusing on user workflow, information presentation, response speed, and the overall graphic design of the interface. The new interface and the underlying distillation capabilities were well-received by analysts during the Utility Evaluation. Quantitatively, the AGILE system was ranked first in many of the evaluation categories, including Overall Usability, Overall Trust, and Overall Usefulness. This user interface was also used to demonstrate GALE distillation capabilities at DARPATech 2007.

Core Technology Improvements

There were many areas of significant technical improvement in Phase 2. Some of the most valuable were:

- Cross-Document Co-Reference. The Phase 1 cross-document algorithm was based primarily on edit distance. During Phase 2 we developed new algorithms involving a wide variety of features, including world knowledge, web knowledge, similarity features, and statistical extraction features. This resulted in fewer overall errors, reducing both spurious name co-reference and missed name co-reference, as well as providing a flexible framework for future algorithm development.
- Answer Selection. In addition to new answer selection patterns, we added new backoff strategies, new pattern score combination techniques, and new ways for document-level characteristics to influence sentence-level answer selection. We began work on techniques to automatically extract complex answer selection patterns from annotation.
- Redundancy Identification and Removal. The redundancy task for Phase 2 was significantly more complex than in Phase 1. Our Phase 2 system identifies redundant "nuggets" by building up similarity from words to nodes in the document parse/proposition trees, combining proposition and paraphrase similarity. Work here focused on exploring mapping techniques and methods for combining all available information to make effective decisions about redundancy, response rankings, and appropriate response lengths.

Go/No-Go Evaluation Performance

On average across the evaluation set, the AGILE system ranked first in almost all of the primary categories of the Go/No-Go Evaluation, including Information Recall, Information Precision, Information F, Document Recall, Document Precision, Document F, and GNG Scaled F-Value. AGILE recall was better than an average time-limited human's, with precision around 60% of an average human's. The AGILE system also performed well on the 24-hour-turn-around "surprise" query, achieving 60% of an average human's performance overall and performing almost three times as well as the closest system competitor.

5.2 Information Sciences Institute (USC)

In the past year, we focused on two goals:

- Investigate the possibility of automated distillation evaluation
- Investigate various aspects of producing fluent and well-formed distillation output

Toward Automated Distillation Evaluation

Since evaluation of distillation engine output is extremely complex and expensive, we investigated the possibility of automatically creating nuggets from text. The nuggets created from gold standard and system output material can then be automatically or semi-automatically compared and the overlap score converted into a Utility goodness score. To be determined is the correlation between this score and the Utility score produced by BAE. (At the first GALE PI meeting of 2007, we obtained verbal support from BAE and LDC, and interest from the GALE community.)

Since the definition of nuggets is somewhat fluid, our approach is to automatically produce minimal-sized nuggets we call eNugs, which can be composed into BAE-sized

nuggets, and which can be compared at the atomic level (thereby enabling the computation of partial overlap credit). We have almost completed the first half of this work. We have defined eNugs in conformance with BAE's nuggets. We have developed code to convert output from either the Charniak parser or the BBN parser Serif into a standard intermediate form. Using Tregex, a regular expression language for trees developed at Stanford, we have developed a set of rules that traverse parse trees and output excerpts of them, which are converted into either nuggets or eNugs. In route to automated evaluation, which requires comparison of eNugs to one another, we tested the feasibility of the approach in a gold standard creation exercise in which two humans used the automatically created eNugs to identify valuable (score-worthy) portions of the texts. Using an interface designed for the task, we achieved satisfactory inter-human agreements.

The postdoc performing this work has left ISI. We are waiting to hear about continued funding in order to hire a replacement and continue with the evaluation portion of the work. The planned work is:

- Produce eNugs from BAE's gold standard material for the 2007 Utility evaluation
- Produce eNugs from BBN's Brandy system output for the 2007 Utility evaluation
- Develop an automated comparison system that matches sets of eNugs (and their near-matches and paraphrases, using a paraphrase table built under this funding last year)
- Experiment with parameter tuning in order to obtain maximal correspondence between system overlap score and BAE's official human-produced Utility scores
- If acceptable, create a software service or package of the evaluation system
- Distribute it to the GALE community.

Producing Fluent and Well-formed Distillation Output

The ISI portion of this work focuses on the problem of redundancy removal, information compaction, and preparation for display, using text summarization and other techniques.

In the early part of Year 2, we built, tested, and delivered to BBN a system to create thumbnail biographies for its distillation system Brandy, to be used in the Utility evaluation only. Before run time, our system created thumbnail bios of each person ever mentioned in the GALE corpus. At run time, while Brandy performed retrieval, extraction, compaction, etc., thumbnail bios of all persons relevant to the query were retrieved and displayed.

Currently, at the request of BBN, we are developing techniques of redundancy removal and compaction, in service of a module that produces coherent and well-formed fluent output for the Brandy distillation engine. As input we accept a ranked set of text snippets, which may be partially redundant. Our approach is to decompose these snippets into smaller pieces, locate and remove redundancies, and reassemble non-redundant material into grammatical and coherent sentence(s), ready for output. This work has recently started and there are no results to report yet.

5.3 Language Computer Corporation (LCC)

Development work for LCC focused mainly on three activities

- preparation for Utility Evaluation (which took place in February 2007)
- preparation for go/no-go Evaluation (which took place in May 2007)
- post- go/no-go evaluation system improvements

The activity for the Utility Evaluation was focused on consolidating PowerAnswer, focusing on: (1) reliability, (2) performance and (3) scalability for multi-user access. In addition, we improved the quality of answers for open-ended questions (non-template-based questions).

The reliability was improved by (a) bug fixes and by (b) providing a level of redundancy in the system.

The performance was improved by:

1. reconfiguring the system, to reduce the I/O effort, by replicating data across the servers (minimizing NFS effort)
2. tuning the system, to find a better balance between computation time and answer quality
3. distributing more computation, using a distributed programming framework, response time was improved.

The scalability for multi-user access was improved by refactoring the system to work reliably for concurrent queries.

The activity for go/no-go evaluation was centered on: (1) providing support for the new templates defined for Phase 2, (2) improving quality of PowerAnswer using technologies that were applied in Phase 1, (3) implementing new technologies for Phase II, and 4) improved corpus processing.

We implemented for Phase 2 templates #2, #12 and #17, using the query reformulation framework that was developed during Phase 1. We implemented gradual improvements to PowerAnswer to improve the output quality for templates both used only in Phase 1 and also for the new Phase 2 templates. We designed and implemented new ways to index and search the document collection, focused mostly for better precision but also for better recall. We designed and integrated report context indexing, to optimize template #5 – we use a report context detection module to identify tuples of "entity who said or reported" and "information that was said or reported" and index them appropriately, to be queried independently during template question answering.

In data processing and in post-processing, we integrated: event detection and indexing, designed to offer (a) better activity date detection and (b) more accurate snippet detection (that matches better the semantics of the template), clause detection, to identify the smallest snippet that embeds the relevant information, proper names disambiguation, to filter inconsistent references to named entities. We enhanced the identification of relevant information in the snippet (e.g.: location and time). The data processing framework was

adapted to an on-demand reindexing, to avoid the prohibitive processing time requirement.

Related to corpus processing, we applied heuristics to do some clean-up of the noisy web data.

The activity post- go/no-go evaluation consisted in: (1) further improvements in the data-cleanup algorithms that are used to improve the Web part of the GALE corpus, (2) prepare the version of PowerAnswer used for the Phase 2 go/no-go evaluation for online access (improved response time), and (3) improvements for bibliographical questions.

6 Accomplishments in Integration and Operational Engines

6.1 BBN Technologies

STT Integration

The primary goal for the AGILE Operational Engines and Integration Task for Year 2 of the GALE program was to identify and transition research algorithms and models, responsible for significant performance gains in STT and MT, into the AGILE operational engine. An additional goal was to define a framework to improve name translation accuracy. BBN achieved significant progress against both these objectives during the course of last year.

BBN completed integration of the AGILE STT research algorithms from the first year of the GALE program (GALE-06) into the AGILE operational engine. The operational engine software base contained numerous engineering refinements from many years of commercial development that were specifically implemented for operational efficiency, speed, and stability. For this reason, we re-implemented each of the new research algorithms inside the operational code base. Some of these algorithms required changes in approach to make them faster, smaller, and incremental with low latency in order to achieve operational viability. Following the successful integration, research STT models can now be directly used in the AGILE operational engine.

The BBN Audio Monitoring Component v2.1 (Year-1 AGILE operational engine) was the baseline system for comparing the performance of the Year-2 AGILE operational engine with GALE-06 STT improvements. All systems were run at 1xRT throughput using standard COTS configuration and compared on the GALE-06 Broadcast News and Broadcast Conversation tuning and development test sets. For the Chinese operational engine, we achieved an impressive 25% relative reduction in character error rate (CER) without incurring an increase in memory footprint. The most dramatic improvement was observed for the conversational test data where we achieved a 41% relative reduction in CER compared to the baseline system. For Arabic, we were able to run both the research Phonetic and Grapheme models in the new operational engine. Using a Grapheme model with a 130K dictionary, we were able to achieve a 27% relative reduction in WER compared to our baseline operational engine with an almost negligible increase in memory size. Both the Arabic and the Chinese operational engines were ready for COTS integration at the end of the year.

Named Entity Recognition and Translation

Accurate translation of names is critically important to the distillation goals of the GALE program. This effort was focused on improving the accuracy of the translation of names as a specific goal that accelerates the critically important subordinate problem of name translation while serving the overall MT goals of the program at the same time. BBN worked closely with Language Weaver (LW) to make considerable progress in three important areas: development of a well-defined and unambiguous ground truth annotation guideline, definition and ground-truth annotation of a development test data set, and definition of evaluation metrics to measure performance that will help guide the research.

Early in the year, BBN and LW proposed an initial approach for ground truth annotation that allowed all reasonable name translations given the context for the name. Initial experiments with this approach exposed serious deficiencies since all plausible name spellings are difficult to specify when multiple variations of the spellings exist for a multi-token name and the variations are combinatorial. BBN and LW proposed two new approaches to counter this deficiency and to ensure high levels of consistency in the ground truth: BBN proposed a rule-driven approach modeled after the practice observed in some of the intelligence agencies, while LW proposed a data-driven guideline style that is acceptable to a US audience and can be derived purely in a data-driven fashion from large corpora. While both approaches have their merits and shortcomings, we believe that the nature of the problem makes it difficult to achieve consensus on one definition of the initial standard across the research community. Further experiments with the annotation guidelines will expose more strengths and weaknesses in the two standards and it is our hope to leverage both standards to define a more consistent ground truth.

BBN selected the Tune/Test/Validation sets defined by the AGILE MT team for 2007 for Name Recognition and Translation research. We used the latest ACE annotation guidelines (Arabic v5.3.3, Chinese v5.5 – May 2005) to annotate five ACE name types, PER, LOC, ORG, GPE, FAC, and one additional type GPE. Nominal in the source transcriptions. During the year, BBN also annotated name translations in a major portion of the test data according to BBN's rule-based annotation guidelines. BBN and LW presented three separate metrics for measuring the performance of Name Recognition and Translation (NRT). In order, these metrics impose increasing levels of constraints on the system output and accordingly require an increasing degree of name recognition capability. Initial experiments were underway by the end of the year to establish baseline benchmarks for NRT research using the annotated ground-truth reference and the specified evaluation metrics.

6.2 Language Weaver (LW)

LW Chinese Syntax-based System Productization

We productized the Chinese-English Syntax-based machine translation system that was developed by ISI and LW. We created a technology platform capable of supporting both phrase-based and syntax-based translation products. We significantly improved the syntax-based translation speed from one word per minute to 650 words per minute via aggressive pruning and beam setting. The final improvements in BLEU over the phrase-based system on an internal test set are summarized in the table below for a variety of

deployment scenarios that range from 4G to 8G of RAM.

System	BLEU (%)	Speed	Memory
LW Chi-Eng Phrase-based	34.48	1098wpm	8G
LW Chi-Eng Syntax-based I	38.62	650wpm	4G
LW Chi-Eng Syntax-based II	41.60	77wpm	8G

LW Arabic-English Phrase-based System Improvements

We significantly improved the LW Arabic-English phrase-based system. We improved our translation and language models by scaling up to training on larger amounts of data. The improvements on the AGILE internal test set are summarized in the table below:

System	BLEU (%)
LW GALE Year I	45.58
LW GALE Year II	50.80

Named-Entity Character-based Transliteration

We devised a character-based statistical approach for transliterating Arabic Named-Entities. We collected Named-Entity transliteration pairs from parallel sentences using rule-based heuristics. We developed algorithms that learn translation and language models at the character level. We implemented a decoder that transliterates an Arabic Named-Entity using the character level models in a log-linear framework with the use of multiple feature functions. We evaluated this approach on Out-Of-Vocabulary (OOV) Arabic Named-Entities. While our generic system fails at translating any of the OOV NEs, our new approach transliterates 38% of the NE OOVs correctly.

Named-Entity Translation/Transliteration Guidelines

We revisited the NE Translation/Transliteration guidelines we developed previously with BBN to better suit them for the general MT Task. We defined a list of guidelines that describe the steps to provide a standard translation/transliteration of a Named-Entity. We started with an initial set of guidelines and then we did three evaluation rounds with three different people where each person was asked to translate, in context, 30 Named-Entities by following the guidelines. The three translations were then analyzed to understand the reasons behind the disagreement. The guidelines were then changed accordingly. The final guidelines will be used to generate a reference dev/test sets for the NE machine translation/transliteration task.

7 Accomplishments in OntoNotes

7.1 BBN Technologies

Assembled and Released the OntoNotes Year 1 Corpus

The first year's worth of OntoNotes training data was released to the LDC in February for distribution first to other GALE sites and then in general. This release included newswire data in English (300K WSJ) and Chinese (250K). An integrated database format was developed, which combines all of the different layers of annotation in a single SQL database. This allowed us to catch and resolve many inconsistencies, as well as supporting model features that combine information from multiple layers. (A paper describing this integrated representation won the Best Paper award at the International Conference on Semantic Computing, ICSC 2007.)

Production Annotation on Year 2 Data

Annotation at all levels proceeded on the Year 2 corpus of broadcast news (200K English and 300K Chinese) and newswire (100K Arabic) data. Improvements in the pipeline and tool support improved the annotation rate and productivity for the word sense annotation.

Preparation of Data for Year 3

We selected broadcast conversation data for our Year 3 corpus, and worked on the Treebank annotation in preparation for doing the other layers of annotation next year.

Data to Support MT

We distributed to GALE sites in November a prerelease of the Treebank for the Chinese broadcast news data from our Year 2 corpus. We also worked with the LDC to arrange for translation of a 50K word portion of the English broadcast conversation data from our Year 3 corpus into Chinese, to increase the amount of parallel data.

Ontology Annotation Path Implemented

Given the increasing number of words for which the word sense annotation has been completed, we were able to design and implement the annotation path that links those senses into the ontology. Annotators form "sense pools" of word senses that are semantically similar, characterizing them with sets of descriptive features, which are then linked to nodes in the upper ontology. We also extended our unified database system to input and maintain the ontology information.

Decoders for Coreference, Propositions, and Word Sense Developed

An initial baseline decoder for coreference was developed, which scored 63.2% F. The new coreference decoder is not limited to a particular set of target types. We did extensive error analysis, comparing our OntoNotes coreference model with the behavior of the ACE Serif system. Performance on types of entities not covered in ACE and on event coreference improved when training on our OntoNotes annotation, and we also improved coreference performance a few points by adding features based on our semantic role label predictions. (A paper on this work was published at ICSC 2007.)

Integration Work with the AGILE Distillation Team

We implemented a combined system that enables our OntoNotes models to be called from within the Distillation system, and began tests to see where OntoNotes features can benefit Distillation performance.

7.2 Information Sciences Institute (USC)

The goal of OntoNotes is to construct by manual work a semantically annotated corpus of several hundred thousand words of English, Chinese, and Arabic text, which has the potential to revolutionize Human Language Technology by making available for the first time semantic information on a large scale. The work at ISI focuses on two aspects:

- Creation of senses for nouns and their annotation in the corpora,
- Pooling of noun senses to form concepts, and insertion into ISI's Omega ontology.

Creation of Senses for Nouns and their Annotation in the Corpora

Noun Sense Creation

This work has gone extremely well. For English we have senses for about 900 English nouns in total, about 400 Chinese nouns, and about 50 Arabic nouns (Chinese started during the past year and Arabic a few months ago).

Noun Sense Annotation

We have double-annotated over 237,000 instances of English nouns (some 800 noun types), mostly from the Wall Street Journal (Year 1) corpus, but also the English-Chinese Treebank and Broadcast News corpus. Of these, over 68,000 achieved 100% inter-annotator agreement. We delivered all the results to BBN in September. Regarding Chinese, we have double-annotated and adjudicated some 21,000 instances (some 170 noun types), which were also delivered to BBN. Regarding Arabic, we have annotated a few hundred instances. Overall, this work has required the services of about 50 annotators, working part-time.

Active Learning

A postdoc visitor on sabbatical to ISI built an Active Learning system that we trained on the Year 1 corpus and deployed on the Year 2 corpus in order to speed up annotation of those nouns that had already been annotated on the Year 1 corpus. Given the difference in sense distributions between the two corpora, we experimented with various mixtures of old and new corpora in order to obtain reasonable performance. The Active Learning system allowed us to save one annotator for about 50% of the high-frequency nouns of the Year 2 corpus. A paper on this work was presented at the EMNLP conference.

Annotation Infrastructure

We have completed all components of the annotation infrastructure. Not counting the STAMP annotation interface that was built at Penn, this comprises four interfaces: the Master Task Handler that displays the words to be annotated and allows annotators to select words and automatically obtain their sentences; the Status interface that displays the agreement results of annotation, for each annotator; the Admin interface that display

the number of hours worked, amount of work done, and various averages for each annotator; a display of the annotation results in detail, allowing computation of correlations, etc. In addition, there is a great deal of data management, including a dedicated server with SVN version control, to make sure nothing is lost.

Pooling of Noun Senses to Form Concepts and Insertion into Omega Ontology

The above work creates lists of word senses. However, only after being cross-correlated do these senses facilitate sophisticated Distillation (for example, by enabling recognition of near-synonyms, cross-part of speech correspondences, etc.). The OntoNotes verb and noun senses are therefore inserted into the Omega ontology. We are building Omega 5, a new version of ISI's ontology, out of the word senses.

Upper Model

To date, we have completed the first version of the Upper Model of Omega 5. This comprises several dozen very high-level concepts for objects and about two dozen for processes (verbs), and provides a fringe of attachment concepts to which the senses (and pools of equivalent senses) can be subordinated. For the Upper Model we also created five 'operator' relations that support multiple perspectives on sense pools / concepts (for example, the Role operator allows the concept Teacher, which is subordinated to Person, to be linked also to Educator).

Pool Creation

We have developed a process that groups equivalent noun (or verb) senses into so-called 'sense pools', which, once ontologized into Omega, function as concepts. This process proceeds in tandem with sense definition. To date, we have created over 1100 sense pools for nouns. These have been delivered to BBN.

Pool Verification

Analogously to the way sense annotation validates the choice and definition of verb and noun senses, we have developed and tested a procedure to validate the pooling of senses. This procedure also requires several validators independently to make decisions over a set of sentences (showing various pool sense alternatives), and requires a high enough level of inter-validator agreement for acceptability. We have determined the appropriate measures and cutoff values. A paper on this work will be delivered at the OntoLex workshop in January.

Pool Insertion

Of the 1062 noun-derived pools available in August, we selected the 300 pools relating to the high-impact Distillation noun set developed by BBN and ontologized them under the Upper Model. The results were delivered to BBN for testing in the Distillation process.

7.3 University of Colorado (CU)

CU is responsible for 10 separate annotation tasks for OntoNotes: (1) English verb sense inventory creation and tagging, (2) pooling and ontology insertion, (3) English PropBanking, (4) English noun predicate annotation, (5) English Treebank/PropBank merge, as well as (6) Chinese treebanking, (7) PropBanking (including noun predicates), (8) Chinese sense inventory creation and tagging, and (9) Arabic sense inventory creation

and tagging and (10) Arabic PropBanking. We have successfully ported all of these processes from Penn to CU. We are on target or close to on target with our Year 2 goals for all of these annotation efforts except English noun predicates due to a delay in receiving the list of nouns from ISI. We are also behind on PropBanking the 300K Chinese Broadcast News, which involves the creation of many new frame files for new verbs.

English Verb Sense Inventory Creation and Tagging

An ITA of 87% or higher is still being achieved for all of the newly trained English sense taggers. There is a high turnover, so we are continually training new taggers. 700 Year 2 verbs have been grouped, 510 have gone through sample annotation, and 240 are tagged and adjudicated. The double annotation for the verbs that passed sample annotation has been delivered. The rest is on target for delivery soon. 4624 senses of 1376 verb lemmas have been inserted into the Omega ontology. This comprises 411 verb sense pools covering 8 semantic classes.

English PropBanking and English Treebank/PropBank Merge

The 200K English Broadcast News was double annotated, adjudicated, and delivered on time. The preliminary evaluation of semantic role labeling done on the revised Treebank/PropBank merge did not show significant improvements.

Chinese Sense Tagging, Treebanking and PropBanking

The 150K Chinese Broadcast Conversation corpus has been TreeBanked. The Chinese verb sense tagging of Year 1 words on the new Broadcast News data has been delivered. The annotation of the Year 2 words on both old and new data is nearing completion. Single annotation of 90% of the 300K Chinese Broadcast News PropBank has been delivered, with the remaining new vocabulary items (10% of the tokens) to be finished shortly soon.

Arabic Sense Tagging and PropBanking

The 454 Arabic PropBank frame files have been examined and 160 of the verb entries have been subdivided into more fine-grained entries for sense tagging. 120 verbs have been tagged, and the next 40 should be annotated be the end of October. We are postponing additional quality control of the Arabic PropBank until the revisions to the Arabic Treebank are complete.

Performance Improvements

We have achieved significant performance improvements in both automatic word sense disambiguation and semantic role labeling for English. For word sense disambiguation we chose just over 200 of the most frequently occurring verbs with adjudicated data and trained our MaxEnt Word Sense Disambiguation system with advanced linguistic features, as well an SVM system. This involved 35K instances. The average ITA for this data was 82%, and both systems also achieved 82% performance. This compares to a performance of approximately 65% on comparable verbs from WordNet.

Portability of statistical parsers and semantic role labeling systems is a major issue, since performance drops some 13% when trained on the WSJ and tested on Brown. We improved the performance of our English Semantic Role Labeling system on Brown

corpus data by 10% (WSJ, 6%) on Arg2 by mapping the PropBank WSJ Arg2 labels to more fine-grained VerbNet semantic roles for training purposes. This created subsets of Arg2's that are more syntactically and semantically coherent, and more robust across genres.

7.4 University of Pennsylvania (Penn)

Syntactic Annotation of 200K words of English Broadcast News Material

Penn hand annotated a 200K corpus of English Broadcast Conversation with syntactic structure after working with LDC to select that corpus. After discussion with the GALE Data Committee, the parallel corpus consisted of 150K words that originated from English Broadcast Conversation (50K of which was translated into Chinese), and 50K of English materials translated from Chinese Broadcast Conversation.

Conversion of Additional 400K words of WSJ material into OntoNotes Syntactic Form

Penn ported an additional 400K words of the Penn WSJ Treebank into the new format for the merged Treebank/Propbank that forms the backbone of OntoNotes, providing a total of 600K words of WSJ for parsing training in the new format.

Participation in Arabic Treebank Advisory Committee

Penn's success late last year in improving Arabic MSA parsing performance from 73.1 F-measure to 79.2 F-measure, largely by changing the POS tag set and parsing structures, played a significant role in triggering the formation of the Arabic Treebank Advisory Committee, on which two Penn AGILE participants sit.

Automatic Case Assignment for Arabic

Systems were developed for automatic case assignment for MSA that assigns case to gold standard ATB trees (hand corrected to yield very low case assignment error) that operates at .9% error (Columbia's rule based approach) to 1.2% error (Penn machine learning result). Penn will move this to automatic parser output as the ATB is cleaned up.

Null Element Placement in Automatic Parser Output for English and Arabic

A global inference model was developed for the key categories of null elements in syntax trees using Conditional Random Fields (CRF) that reduced error rates on determining the underlying role of Nominal Wh-phrases ("who," "what, etc.) from our earlier state of the art result of 11.6% using a cascade of independent linear classifiers to 5.4%. Initial application of this new null-element inference algorithm to gold-standard Arabic ATB parses yields 70% correct performance on nominal wh-traces, 75% on adverbial, and 85% on topicalization, providing a baseline for improvement during Year 3.

OntoNotes Viewer

Initial development of a GUI based Viewer for the unified OntoNotes relational database has been completed, with an initial release as part of the OntoNotes 2.0 release. OntoBrowser allows simultaneous browsing of (a) Treebanks, including expansion and compression of individual parse nodes, in both tree and Penn Treebank bracketed formats, (b) PropBanks, (c) Coreference structures, and (d) Word senses. OntoBrowser uses a variety of graphical devices to show the linkages between these various structures.

8 Technology Transition

Numerous transition efforts during GALE Year 2 resulted in the deployment and subsequent operational use of GALE Year 1 technology. While some of these efforts were funded by government sponsors including DARPA and TSWG, others were funded directly by end users. This indicates a strong desire by operational users to integrate the advanced human language technology developed under the GALE program into their workflow.

In December 2006, a two-channel system of BBN Broadcast Monitoring System BMS v2.0 was deployed in Balad, Iraq, for the 5th Special Forces Group. This system contains the operational improvements made to the AGILE STT and MT engine components under the GALE Year 1 effort. The deployment was supported by the DARPA Director's office. At the request of DARPA, BBN assembled and tested the system very rapidly and shipped it to Iraq within 21 days after receipt of the Purchase Order. In parallel, BBN's technical staff trained a subcontractor to install the system in the field and to train the users and the system administrators. The subcontractor traveled to Iraq and successfully installed the system in December 2006. The system was installed within three hours and has been operational since then. The subcontractor trained over 15 users and demonstrated the system and its utility to several senior officers and visitors during his 10-day stay at the site. Since the installation, BBN has regularly followed up with the designated system administrators for this equipment.

A four-channel Chinese BMS v2.0 was deployed to a location in the Far East during December 2006. This deployment was supported by the customer and coordinated by a large Prime Contractor working for the customer. BBN is a subcontractor on this effort. More deployments of the BMS are anticipated in 2007 to serve this customer's world-wide OSINT collection and analysis operations.

In December 2006, A two-channel BBN BMS v2.0 was installed in Baghdad, Iraq, for the Strategic Effects Directorate of the Multi-National Force-Iraq (MNF-I). This system deployment was supported by the Technical Support Working Group (TSWG).

In February 2007, a single-channel Arabic BMS v2.0 was deployed to the DCGS-A Systems Integration Laboratory at Ft. Monmouth, NJ. This deployment was supported by the U.S. Army CECOM and was facilitated by the Sequoyah Transition Management Office (S-TMO).

During spring and summer of 2007, BBN developed Farsi/Persian models for the BMS. Although Farsi is not funded by the GALE program, this effort, funded by TSWG, was able to leverage many of the research developments produced by GALE. The addition of Farsi increases the number of languages supported by BMS to four: Modern Standard Arabic, Mandarin Chinese, Western Hemisphere Spanish, and Farsi. Farsi was made commercially available in September 2007.

During the spring of 2007, BBN negotiated a subcontract agreement with SOS International, Inc. (SOSi) for Operations and Maintenance services to support our OCONUS deployments of the BBN BMS in Iraq. These systems include the commercial version of the AGILE Operational Engine. In June, a SOSi employee, who had been previously trained at BBN's facility in Cambridge, MA, visited the 10th Special Forces

Group at Camp Anaconda, Balad, Iraq. The SOSi contractor remained on site in Balad for 4 days, providing troubleshooting and user training services for the BMS that was deployed to this site in December 2006 for DARPA.

In September, 2007, BBN upgraded a two-channel Arabic BMS used by the SOJICC program at SOCOM HQ in Tampa, Florida, from version 1.3 to 2.0. This system was used operationally by SOCOM personnel, and funding for this upgrade was provided by TSWG.

In addition, SOCOM notified BBN of its intention to expand its operational BMS installation to a total of five channels – two for Arabic, and one each for Mandarin Chinese, Western Hemisphere Spanish, and Farsi. SOCOM has also paid for annual maintenance services from BBN since July 2006. This early sale of Farsi, accompanied by these repeated commitments by SOCOM, demonstrate a critical level of user acceptance of the technology in meeting their operational mission requirements. This transition represents a successful maturation of a short-term technology insertion (supported by R&D funds) into a long-term material and maintenance procurement (supported by the end user) to serve an expanding operational mission.

9 Infrastructure Contributions to GALE Program

BBN

- Provided a sizable amount of Chinese text (~300M characters) to be used for language modeling for Mandarin STT
- Language model training corpus LDC2007E24 consisting of news articles downloaded from the web (about 1B words)
- 5B words of English web text
- Release of OntoNotes 1.0; 300K English and 250K Chinese newswire annotated with syntactic structure, propositions, word sense, and coreference, packaged in an integrated relational database
- Ground truth test sets to support name translation research
    - Arabic and Chinese; development, tuning, and evaluation MT-06 test sets (6 total)
    - Name types defined by ACE guidelines; name translation guidelines under development
    - will be given to LDC before PI meeting
- Scoring scripts and GLM for Arabic.
- Provided LDC with tools for measuring document perplexity to assist in data selection for the evaluation
- External AGILE STT/MT servers set up at BBN for daily use by LDC over the Internet
    - Arabic and Chinese; transcription and translation engines (4 total)
    - Custom client software created for LDC to handle network interactions

- Set up 4 external servers that ran AGILE year-1 operational engines for Arabic and Chinese to support LDC data collection and annotation work. There are two audio servers capable of both STT and MT, and two text servers that are dedicated to MT.
- Provided Lincoln Labs with four copies of a pre-release of the AGILE year-2 operational engine software, two Arabic systems and two Chinese systems for an onsite study for GALE.
- Continued to support the design and development of UIMA and the third version of the Inter-Operability Demo (IOD3).

Cambridge University

- Language Model Training Data
  - Arabic web data (300M words, 24 sources)
  - Mandarin web data (400M characters, 18 sources)
- Evaluation/Development set
  - Mandarin STT Dev07 specification and initial references
  - Mandarin STT Eval06/Eval07 reference processing/refinement
  - Corrected STT reference for the Arabic common devset dev07
- Evaluation/Development set
  - Mandarin STT Dev07 specification and initial references
  - Mandarin STT Eval06/Eval07 reference processing/refinement
  - Corrected STT reference for the Arabic common devset dev07
- HTK Version 3.4 software release (Dec. 2006); includes
  - discriminative training (MPE/MMI)
  - large vocabulary decoder (HDecode)
- MTTK Parallel Text Alignment Modeling Toolkit
  - public domain release of 2006 continues to be downloaded
  - updated release anticipated by end of 2007 (tentative) with improved alignment models and rescoring procedures

Sakhr

- Parallel corpus: 19M words Arabic → English; 8M words English → Arabic
- Manually vocalized Arabic acoustic training data (FBIS/BBN/Sakhr corpus)

LIMSI

- Provided downloaded web texts to community
- Word aligner provided to NIST